DOCUMENT RESUME

ED 402 321 TM 025 815

AUTHOR Manhart, Jim J.
TITLE Factor Analytic Methods for Determining Whether Multiple-Choice and Constructed-Response Tests Measure the Same Construct.
PUB DATE Apr 96
NOTE 48p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, NY, April 9-11, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Chi Square; *Constructed Response; Goodness of Fit; High Schools; *High School Students; Models; *Multiple Choice Tests; *Sciences; Test Construction; *Test Items
IDENTIFIERS *Confirmatory Factor Analysis; *Item Parcels; Test Specifications

ABSTRACT

The relationship between multiple-choice and constructed-response science tests was investigated using confirmatory factor analysis. The tests were given to 872 students in grades 9 through 12. Each test was divided into several parcels of items. The fit of a one-factor model (parcels of both tests loading on the same factor) was compared with the fit of a two-factor model (parcels of each test loading on a different factor). Inspection of chi-square data and standardized residuals led to the conclusion that the two-factor model was generally more appropriate for explaining the covariance between the parcels than the one-factor model. The conclusion about the number of factors was consistent across six methods that varied the number of parcels used and how items were assigned to the parcels. Two appendixes present the test specifications for the multiple-choice and constructed-response tests. (Contains 8 tables and 16 references.) (Author/SLD)

FACTOR ANALYTIC METHODS FOR DETERMINING WHETHER MULTIPLE-CHOICE AND CONSTRUCTED-RESPONSE TESTS MEASURE THE SAME CONSTRUCT

Jim J. Manhart

University of Iowa

Paper presented at the 1996 annual meeting of the National Council on Measurement in Education

ABSTRACT

The relationship between multiple-choice and constructed-response science tests was investigated using confirmatory factor analysis. The tests were given to 872 students in grades 9-12. Each test was divided into several parcels of items. The fit of a one-factor model (parcels of both tests loading on the same factor) was compared with the fit of a two-factor model (parcels of each test loading on a different factor). Inspection of chi-square data and standardized residuals led to the conclusion that the two-factor model was generally more appropriate for explaining the covariance between the parcels than the one-factor model. The conclusion about the number of factors was consistent across six methods which varied the number of parcels used and how items were assigned to the parcels.

TABLE OF CONTENTS

INTRODUCTION[1]

The wisdom of relying solely on the multiple-choice format in standardized achievement testing has been questioned. Critics claim that multiple-choice questions are unable to assess some educational objectives, especially objectives involving higher order cognitive skills. If this criticism is legitimate, the validity of interpreting scores on a multiple-choice test as indicating achievement of such objectives may be limited due to construct underrepresentation. Moreover, if teachers gear instruction to what is found on standardized tests, objectives that aren't assessed by multiple-choice items will tend to be eliminated as targets for learning (Frederiksen, 1984).

In response to these criticisms, test publishers have begun to provide constructed-response as well as multiple-choice items. For instance, constructed-response tests have recently been published to supplement the multiple-choice Tests of Achievement and Proficiency (TAP) in English, math, science and social studies. Questions about the relationship between these constructed-response and multiple-choice tests arise. Specifically, do the constructed-response tests measure something different than the corresponding multiple-choice tests?

Research on the relationship between constructed-response and multiple-choice tests has yielded conclusions that appear to be

[1]The author gratefully acknowledges the help of Dr. Robert Forsyth, Dr. Stephen Dunbar and Dr. Robert Ankenmann who read earlier drafts of this paper and provided suggestions for improvement.

inconsistent; some studies suggest that the two types of tests measure constructs that are psychometrically equivalent, while other studies indicate the two types of tests measure different constructs. Frederiksen (1984) claimed that when constructed-response questions are nothing more than multiple-choice stems with the choices omitted, little difference is typically found between the two types of tests. On the other hand, when Ward, Frederiksen and Carlson (1980) used higher order constructed-response items to create multiple-choice items (rather than the reverse), the two tests appeared to measure different constructs. Ackerman and Smith (1988) concluded from their review of the literature that item format has little effect on the construct being assessed, the Ward, Frederiksen & Carlson (1980) study being an exception to the rule. However, their own study on writing assessment demonstrated that there was a difference between constructed-response and multiple-choice tests (Ackerman & Smith, 1988). Part of the reason for inconsistent conclusions may stem from failing to carefully distinguish between tests that differ only in item format, and tests that differ not only in item format but also in the cognitive skills required of the examinee.

Whether constructed-response and multiple-choice tests measure different constructs may also be dependent on the content domain. Traub (1993) tentatively concluded that with respect to the writing domain, the different item formats measure different constructs, whereas for the reading comprehension and quantitative domains the item format appears to make little difference. However, this

conclusion was based on a review consisting of only nine studies, and thus there is a need for further research on the relationship between multiple-choice and constructed-response tests. It should also be noted that even if the two types of tests are found to assess psychometrically equivalent constructs, it does not follow that they necessarily provide the same information. For instance, constructed-response tests may provide diagnostic information that the multiple-choice tests do not (Birenbaum & Tatsuoka, 1987).

In investigating the relationship between multiple-choice and constructed-response tests, confirmatory factor analysis can prove useful. If each type of test is divided into several parcels, confirmatory factor analysis can be helpful in deciding whether a two-factor model (each type of test loading on its own factor) or a one-factor model (both tests loading on the same factor) is more appropriate for explaining the covariance of the parcels. If the two-factor model is found to be more appropriate, this conclusion in conjunction with other evidence such as a content analysis of the two tests, could be used to build an argument that the two tests measure different constructs. On the other hand, if a one-factor model is more appropriate, the two tests may be treated as equivalent measures from a psychometric standpoint, although a content analysis of the tests might still suggest that they are measuring different constructs which happen to be correlated. Thus, confirmatory factor analysis can provide information about the relationship between multiple-choice and constructed-response tests, but by itself, is insufficient for determining whether the

two types of tests measure different constructs.

The present investigation was designed to examine the robustness of this confirmatory factor analysis method. The multiple-choice and constructed-response TAP science tests were parceled, and factor analytic methods were used to decide whether a one-factor model or a two-factor model is more appropriate for explaining the covariance of the parcels. Of particular interest was whether the conclusion about the number of factors is independent of the number of parcels used and how items are assigned to the parcels.

BACKGROUND

A traditional method for investigating the relationship between two tests employs a result from classical test theory, namely, the correction for attenuation formula. According to classical test theory, the observed score on a test is equal to the sum of the true score and the error component. If two tests measure the same construct, the correlation between their true scores is 1.0. However, true scores are hypothetical and thus can not be directly observed. The correlation between the observed scores of the two tests will be less than 1.0 due to random measurement error. The correction for attenuation formula estimates what the correlation between the observed scores would be if there had been no measurement error, that is, it estimates the correlation between the true scores. The closer the correction for attenuation is to 1.0, the stronger the evidence that the rank-order of the examinees' true scores is the same for both tests, and thus the two tests are measuring psychometrically equivalent constructs. Since this corrected correlation does not have the same sampling distribution as other correlations, a special hypothesis test is needed to determine whether the disattenuated correlation differs significantly from 1.0 (Lord, 1957). In actual use, the correction for attenuation sometimes leads to values greater than 1.0 as well as other problems, and so it should be used with caution, if at all (Winne & Belfry, 1982).

An alternate way for investigating the relationship between two tests employs confirmatory factor analysis. One method using

confirmatory factor analysis involves first dividing each test into subtests or parcels. The data are fit with a two-factor model in which the parcels of one test are restricted to load only on the first factor and the parcels of the other test are restricted to load only on the second factor, but the factors are allowed to be correlated. A comparison is then made between this model and a model where all parcels are restricted to load on a single factor. This comparison between the two models generally involves looking at various indicators of how well each model fits the observed covariance matrix of the parcels. Indicators of fit include the chi-square values and standardized residuals for each model. Practical significance should also be taken into account. This is especially necessary since the chi-square tests depend on sample size and thus with a large sample, small differences are likely to be detected that are of little practical significance (Loehlin, 1992, p. 65).

The classical test theory and confirmatory factor analysis methods for investigating the relationship between two tests are not as different as they may at first appear to be. In classical test theory, the observed score variance is equal to the sum of the true score variance and the error variance. The factor analysis model also partitions the observed score variance into two parts, that is, the observed score variance is the sum of the variance explained by a common factor and the variance unique to the particular test. Conceptually, the unique variance is further partitioned as the sum of variance specific to the test and error

variance. If the specific variance can be assumed to be zero (or at least small), then for practical purposes the true score and error variances of classical test theory correspond respectively to the common variance and unique variance of factor analysis.

With the assumption that parcels from the same test have negligible specific variance, the factor analysis method for investigating the relationship between two tests can be reformulated in terms that closely resemble the classical test theory method, namely, as a hypothesis test that the disattenuated correlation coefficient equals 1.0 (Joreskog, 1971). As before, the data are fit with a two-factor model where the parcels of each test are allowed to load on only one factor, although the two factors are allowed to be correlated. However, given the connection between classical test theory and factor analysis outlined in the previous paragraph, the correlation between the two factors is the disattenuated correlation coefficient. If the data are fit again to the two-factor model but with the correlation between the factors fixed at 1.0 (i.e., a one-factor model), the chi-square difference between these two models gives a test that the disattenuated correlation coefficient is equal to 1.0.

This reformulation of the confirmatory factor analysis method is similar to the classical test theory method, but not exactly the same. The classical method requires that the test parcels be two parallel tests, that is, the parcels for a given test would need to have the same true score variance and the same error score variance. Indeed, the classical method could be tested using

confirmatory factor analysis by further restricting the two-factor model so that the factor loadings for a given factor are equal and the unique variances for parcels loading on the factor are equal. Without these restrictions, the parcels are congeneric tests, that is, they do not necessarily have equal true score variances nor equal error variances. The congeneric model includes parallel parcels as a subset, but to use confirmatory factor analysis there is no need to meet the more restrictive parallel test assumptions.

For the present investigation, where one of the tests contains constructed-response items of differing point value, it is very unlikely that the parcels will be classically parallel. However, the less restrictive congeneric assumptions can be met, and thus the confirmatory factor analysis method is favored over the classical test theory method for testing whether the disattenuated correlation equals one. A further advantage of the factor analysis method is that it includes additional indicators of fit (e.g., standardized residuals) rather than relying solely on the disattenuated correlation hypothesis test.

The congeneric confirmatory factor analysis model was employed in a recent study comparing the multiple-choice and constructed-response sections of the Advanced Placement Computer Science test (Bennett, Rock, & Wang, 1991). The fifty multiple-choice questions were divided into five parcels that were stratified with respect to content and had approximately the same average difficulty. Each of the five constructed-response questions served as a parcel. Although the two-factor model had been hypothesized, the

researchers concluded that the one-factor model sufficiently explained the data. Thus, even when a two-factor model is anticipated, the one-factor model may be about as good at explaining the data as the two-factor model.

Consideration of the way the parcels were constructed in the above study raises some potential research questions. First, although it was convenient to divide the multiple-choice test into five parcels (so each parcel had the same number of questions), would the fit indicators change significantly if a different number of parcels had been used? Loehlin (1992, p. 64) suggests that as few as three parcels can be used to adequately mark a factor. Second, is it necessary to stratify the parcels by content and ensure that the parcels have approximately the same average difficulty? Would the results differ significantly if, for example, every 5th item had been assigned to a parcel?

Cook, Dorans and Eignor (1988) also used parcels in a study of the dimensionality of the Sat-Verbal test. As to the number of parcels that can be used, they suggest that as long as each parcel contains 6 or 7 items so that the score distribution of the parcel approximates a normal distribution, it should make little difference how many parcels are used. They also point out that the reason for using parcels rather than the individual items is to help ensure that the covariance matrix is not a function of item difficulty or affected by violations of the linear regression assumption on which the factor analysis model is based. To do this, they claim it is "essential to place approximately equal

numbers of easy, middle difficulty, and hard items within each parcel" (Cook, Dorans & Eignor, 1988, p. 26). However, the question still remains as to whether the results of factor analysis are significantly changed when this requirement that the parcels are of approximately equal difficulty is not met.

RESEARCH QUESTIONS

The present research addressed the following questions:

1. Is a one-factor model or a two-factor model more appropriate for explaining the covariance of parcels marking the TAP multiple-choice and constructed-response science tests?
2. Are the confirmatory factor analysis indicators of fit independent of the number of parcels used and how items are assigned to parcels?

METHODS

Subjects

Subjects consisted of a subset of students who were part of the TAP national standardization sample (Fall, 1992). The subjects were all from the same school district and were in grades 9-12. They consisted of 257 ninth graders, 259 tenth graders, 169 eleventh graders and 187 twelfth graders. Further information about the subjects was not available.

Instruments

The multiple-choice tests were the TAP science tests, Form L, Levels 15-18 (Riverside Publishing, 1993b). These tests consists of 50 items worth one point each. Twenty-five of the multiple-choice questions for each grade also appear on the test for the next grade. Thus, the multiple-choice test for grade 10 consists of 25 questions found on the grade 9 test and 25 questions found on the grade 11 test. The constructed-response tests were the TAP Performance Assessments for science (Riverside Publishing, 1993a). The number of items on these tests ranges from 18 to 26, with the maximum point value of the items ranging from one to four. The constructed-response questions are unique for each grade, that is, a given question is only asked at one grade. Table A summarizes the number of subjects and the instruments used at each grade. Readers interested in the content covered in each test are referred to Appendix A and Appendix B, which contain the test specifications for the multiple-choice tests and the constructed-response tests, respectively.

Procedure

The mean, standard deviation, item difficulties, average difficulty and coefficient alpha were calculated for both the multiple-choice and constructed-response tests. Since items on the constructed-response tests vary in point value, the difficulty for each constructed-response item was calculated by dividing the item mean by the item's maximum point value. The average difficulty for the constructed-response tests was computed as a weighted average, with the items' maximum point value used for the weights.

The multiple-choice tests were divided into parcels in six ways. To begin with, three parcels were created using a content stratified/equal difficulty procedure. This consisted of initially assigning every third item in a given content stratum to one of the parcels. Items were then interchanged between the parcels but within content strata until the average item difficulty was about equal for all three parcels. This content stratified/equal difficulty procedure was repeated to divide the multiple-choice test into two parcels, and repeated again to create five parcels. Three additional procedures were used in assigning items to create three parcels. One procedure assigned every third item on the multiple-choice test to each parcel. Another procedure consisted of placing life science questions in the first parcel, earth science questions in the second parcel and physical science questions in the third parcel. (Note: Items classified under the Nature of Science/Scientific Process category in the TAP test specifications were reclassified based on whether the item's

content was life, earth or physical science, and then assigned to the corresponding parcel). The final procedure assigned the hardest 17 items to the first parcel, the easiest 17 items to the third parcel and the remaining 16 items to the second parcel.

The constructed-response tests were divided into parcels in just one way. The decision to use only one procedure for creating the constructed-response parcels was made because these tests are still very experimental and have far fewer questions to assign to each parcel. Therefore, the constructed-response tests were only divided into three parcels, the minimum number suggested to adequently mark a factor (Loehlin, 1992, p. 64). The items were assigned to the parcels by the content stratified/equal difficulty procedure. However, the content classification of the test specifications for the constructed-response tests does not use traditional subject matter categories. The four content areas are nature of science, science subject matter, scientific concepts/connections and decision making/communication. These nontraditional content classifications were used to stratify the parcels, but content stratifications using more traditional content categories were also taken into account. For example, all the grade 10 test items (biological) were classified more traditionally as dealing with either anatomy or classification. Items were then interchanged between parcels but within content strata until the weighted average difficulty was approximately the same for all three parcels.

A summary of the various parcel methods (PM1-PM6) is provided

in Table B for easy reference. For each of the four grades, the multiple-choice tests were parceled in six different ways while the performance assessments were parceled in only one way, yielding a total of 24 combinations. The covariances and correlations between the parcels were computed for each combination. The covariance matrices were then submitted to LISREL 7 for confirmatory factor analysis using maximum likelihood estimation. Every covariance matrix was fit with a two-factor model (each type of test loading on its own factor) and a one-factor model (both tests loading on the same factor).

Data Analysis

The descriptive statistics for both the multiple-choice and constructed-response tests were examined first. The subjects used in this study were a convenience sample, and so the descriptive statistics for the multiple-choice tests were used to compare the sample of this study to the standardization sample to see if there were any major differences between the two samples. (Such a comparison can not be made for the constructed-response tests because the sample of this study consisted of all the students who took those tests.)

Since part of the rationale behind the content stratified/ equal difficulty method is to produce parcels that are very similar and thus are equivalent markers of the original test, the range of intercorrelations between parcels from the same test was examined. For the methods that created three multiple-choice parcels (PM1, PM4, PM5, PM6), a statistical test was employed to determine

whether the highest and lowest intercorrelations were significantly different (Hinkle, Wiersma & Jurs, 1988, p. 280). This statistical test was also conducted on the three constructed-response parcels. Across the four grades, this amounted to 20 separate statistical tests. To keep the overall chance of a Type I error to a reasonable level, the .01 significance level was used each time.

For each grade, the initial analysis consisted of submitting the covariance matrix of PM1 to confirmatory factor analysis. The one-factor and two-factor models for PM1 are illustrated below and the estimated parameters for each model are itemized.

KEY MC = multiple-choice
CR = constructed-response
P = parcel
x = factor loading estimated
o = factor loading not estimated

ONE-FACTOR MODEL

	1
MC P1	x
MC P2	x
MC P3	x
CR P1	x
CR P2	x
CR P3	x

Estimated Parameters
6 factor loadings
6 error terms

TWO-FACTOR MODEL

	1	2
MC P1	x	o
MC P2	x	o
MC P3	x	o
CR P1	o	x
CR P2	o	x
CR P3	o	x

Estimated Parameters
6 factor loadings
6 error terms
1 correlation between factors

The conclusion as to whether a one-factor model or a two-factor model is more appropriate for explaining the covariance of the parcels was arrived at by inspection of chi-square values, chi-square per degree of freedom, difference in chi-square, standardized residuals, and modification indices. A brief description of how each of these fit indicators was interpreted is

given in Table C. The difference in chi-square values is appropriate to use since the models are clearly nested; if the correlation between factors in the two-factor model is set equal to 1.0, the two-factor model collapses to the one-factor model.

The effect of the number of parcels on the various fit indicators was then investigated. This consisted of comparing PM1, PM2, and PM3. Specifically, the chi-square value per degree of freedom, difference in chi-square, standardized residuals and modification indices were examined to see if there were any consistent trends across the four grades. The effect of how items are assigned to parcels was assessed in a similar manner by comparing PM1, PM4, PM5, and PM6.

RESULTS

Descriptive statistics for both the multiple-choice and constructed-response tests are given in Table D. The statistics shown in Table D for the standardization sample were provided by Riverside Publishing (personal communication, August 19, 1995) since they have not yet been published. For the multiple-choice tests, the descriptive statistics for the sample of this study are relatively similar to those for the standardization sample at each grade. In general, both samples tend to have an average difficulty around .5, a coefficient alpha of about .9 and a standard deviation of about 10 for each grade. For the constructed-response tests, the descriptive statistics reveal differences between grade 9 and the other three grades. The grade 9 constructed-response test was relatively difficult (average difficulty = .3069), while the tests for the other three grades had average difficulties in the .4 to .6 range. The grade 9 constructed-response test also had a much lower standard deviation (2.95) and coefficient alpha (.503) than the corresponding tests for the other grades, which had standard deviations around 5 or 6 and coefficient alphas above .7. The correlation between the multiple-choice and constructed-response tests varied from .537 to .687 across the four grades.

For each of the parcel methods, Table E shows the range of intercorrelations between parcels from the same test. In comparing intercorrelations for multiple-choice parcels formed by the content stratified/equal difficulty procedure, the intercorrelations are highest with two parcels (PM2), intermediate with three parcels

(PM1) and lowest with five parcels (PM3). For those methods resulting in three multiple-choice parcels, PM5 (by content) and PM6 (by difficulty) generally resulted in a statistically significant difference between the highest and the lowest intercorrelations ($p<.01$). The only exception is grade 11, PM5. On the other hand, PM1 (content stratified/equal difficulty) and PM4 (every third item) did not yield a statistically significant difference between the highest and the lowest intercorrelations for multiple-choice parcels. The intercorrelations of the constructed-response parcels (formed by the content stratified/equal difficulty procedure) were also not statistically different.

Results of the initial confirmatory factor analysis which used the covariance matrix of PM1 are shown in Table F. The results are fairly consistent across all four grades. The chi-square value for the one-factor model is significant at the .05 level, whereas the chi-square value for the two-factor model is not. The chi-square per degree of freedom ratio for the one-factor model is over 2, while for the two-factor model it is under 2. The chi-square difference between the two models is statistically significant ($p<.001$). Thus, the chi-square data suggests that the two-factor model is more appropriate than the one-factor model.

More support for the two-factor model over the one-factor model is given by the standardized residuals. With the exception of grade 9, a fair number of the standardized residuals for the one-factor model are $> |2.58|$. In addition, there tended to be a pattern to the negative and positive standardized residuals for the

one-factor model. This pattern is illustrated in the following matrix of standardized residuals (for grade 10, one-factor model, PM1):

MATRIX OF STANDARDIZED RESIDUALS WITH PATTERN
MC = Multiple-choice parcel; CR = Constructed-response parcel

	MC1	MC2	MC3	CR1	CR2	CR3
MC1	0					
MC2	1.724	0				
MC3	1.211	0.544	0			
CR1	-3.041	-0.338	-0.168	0		
CR2	-0.110	-1.075	-1.740	2.585	0	
CR3	-1.093	-1.933	-0.530	3.978	2.806	0

Note that there are positive residuals when fitting the covariances among the multiple-choice parcels and when fitting the covariances among the constructed-response parcels, but negative residuals when fitting the covariances between multiple-choice and constructed-response parcels. When Table F lists a "yes" for the pattern, the matrix of standardized residuals either matched the prototype shown above exactly or contained just one residual deviating from this prototype. Except for grade 9, this pattern was present for the one-factor model. On the other hand, for the two-factor model the pattern of positive and negative residuals disappeared, and none of the standardized residuals are > |2.58|.

Therefore, examination of the chi-square data and standardized residuals generally favors the two-factor model over the one-factor model. Across the four grades, the correlation between the two factors ranges from .672 to .839. The specified two-factor model (each test loading only on it own factor) is further supported in that the maximum modification index is always less than 5. This indicates that allowing any of the parcels to load on both of the

factors would not significantly improve the model, which is consistent with the rationale behind the methodology of this study. However, it should be noted that the two-factor model is not as clearly favored for grade 9 as for the other grades. For grade 9, the chi-square value for the one-factor model was rejected at the .05 level but would not be rejected at the .01 level, the chi-square per degree of freedom ratio for the one-factor model is very close to 2 and the chi-square difference is much lower than for the other grades. Moreover, the standardized residuals do not indicate a significant problem with the one-factor model in that only 1 was > |2.58| and the characteristic pattern was not present. Thus, the grade 9 data is somewhat ambiguous with respect to whether the one-factor model or the two-factor model is more appropriate.

Data relevant for assessing the effect of varying the number of parcels is reported in Table G. Note that changing the number of parcels changes the degrees of freedom, so the overall chi-square value is not directly comparable across PM1, PM2 and PM3 (3, 2 and 5 parcels, respectively). The chi-square per degree of freedom ratio should be used in comparisons. As was true for the initial analysis, the results are fairly consistent across the four grades. Regardless of the number of parcels, the chi-square ratio for the one-factor model was generally over 2, while for the two-factor model it was always under 2. The only exception was the grade 9, one-factor, PM3 ratio of 1.896. The chi-square difference was consistently significant ($p<.001$) across number of parcels. Thus, the conclusion from the chi-square data appears to be

unaffected by varying the number of parcels, that is, the two-factor model is consistently favored over the one-factor model.

In examining the standardized residuals for the methods varying the number of parcels, the one-factor model tends to have a fair number > |2.58|, while the two-factor model has none or only one > |2.58|. The pattern described earlier for the one-factor model using three parcels (PM1) was also observed when two parcels (PM2) were used, but was only observed for grade 11 when five parcels (PM3) were used. Overall, the examination of the standardized residuals suggests the two-factor model is a more appropriate model than the one-factor model, although the characteristic pattern for the one-factor model tended to disappear when five parcels were used. For a given grade, the correlation between the two factors remained fairly constant regardless of the number of parcels. The maximum modification index for PM2 was always under 5, but for PM3 the maximum modification index for one grade was slightly over 5, namely, the 5.230 value for grade 12.

Data for examining the effect of how items are assigned to parcels is found in Table H. Except for grade 9, the chi-square data is consistent across the four assignment methods used in this study. The chi-square value for the one-factor model was significant at the .05 level, while for the two-factor model it was not. The chi-square ratio was over 2 for the one-factor model and under 2 for the two-factor model. The chi-square difference was significant ($p<.001$). Thus, regardless of the assignment method, the chi-square data tends to favor the two-factor model over the

one-factor model. The grade 9 chi-square data violates this general trend for two of the parcel methods. First, the one-factor model for PM5 was not significant at the .05 level and the chi-square ratio was under 2. Second, the two-factor model for PM6 was significant at the .05 level and the chi-square ratio was over 2.

Regardless of how items were assigned to parcels, a fair number of standardized residuals for the one-factor model were > |2.58|, while for the two-factor model none were > |2.58|. Other than grade 9, the only exception to this rule was for grade 12, two-factor model, PM6 where one of the 15 standardized residuals was > |2.58|. The standardized residuals for the one-factor model generally showed the characteristic pattern, while the two-factor model did not. Other than grade 9, the only exception was grade 12, one-factor model, PM6 failed to show the pattern. The grade 9 standardized residuals for the one-factor model > |2.58| tend to be fewer than for the other grades and the grade 9, two-factor model, PM6 still had four standardized residuals > |2.58|. Also for grade 9, the one-factor model does not show the standardized residual pattern. Over all the grades however, the standardized residual data tends to support the two-factor model over the one-factor model regardless of the method used to assign items to parcels. For a given grade, the correlation between the two factors did not vary much across the various methods of assigning items to parcels. The maximum modification index tended to stay under 5, although it was over 5 in three instances (i.e., 9.108 for grade 9, PM6; 5.193 for grade 12, PM5; and 7.839 for grade 12, PM6).

DISCUSSION

The results reported in the previous section are fairly consistent across the grades except for grade 9. The descriptive statistics for the grade 9 constructed-response test indicate a problem. This test was much more difficult than the other constructed-response tests. Because of the difficulty, the variance and the reliability (internal consistency) of this test are much lower than for the constructed-response tests given to the other grades. In light of this problem, the conclusions reached in this discussion are based more on the data from grades 10, 11 and 12 than from grade 9. An examination of the items on the grade 9 constructed-response test (Bicycle Science) revealed that a fair number of the questions deal with the algebraic formulas for such concepts as mechanical advantage, work, power and momentum. Although these are "simple" formulas, the typical ninth grader is just starting to study algebra and thus it's likely that most ninth graders will not have acquired the cognitive skills necessary to manipulate even these simple formulas.

For the initial analysis using PM1, both the chi-square data and the standardized residuals favored the two-factor model over the one-factor model. This result could be used in building an argument that the two types of tests are measuring different constructs. The actual items on the two tests would need to be thoroughly examined in order to name and describe the two constructs. An examination of the items on each test might reveal that the two tests differ in the content they assess and/or differ

in the cognitive skills they require of the test taker. Such an examination is beyond the scope of the research questions addressed by this study and therefore will not be pursued further here.

When the number of parcels was varied, neither the chi-square data nor the standardized residuals were significantly affected. Regardless of the number of parcels, both the chi-square data and the standardized residuals suggested the two-factor model was more appropriate than the one-factor model. However, the characteristic pattern for the one-factor model that was observed using three parcels, was only observed for one of the grades when five parcels were used. Also the maximum modification index for grade 12 when five parcels were used (PM3) was over 5. This indicates a better fitting model would result if one of the parcels were allowed to load on both factors, in contradiction to the rationale behind the methodology of this study. Thus, five parcels may be an upper limit for the number of parcels that can be used. Also, Table E shows the range of intercorrelations for five parcels spans about .1 in the correlation metric, while for three parcels it is less than half this much. Therefore, using three parcels is preferable to using five. Moreover, since to mark a factor three parcels are better than two parcels, three may be the optimal number of parcels to use.

Comparison of the four methods for varying how items were assigned to parcels also showed there was little effect on the chi-square data and standardized residuals. Regardless of how items were assigned to parcels, the two-factor model was still favored

over the one-factor model. On the other hand, the data in Table E showing the range of intercorrelations suggests that the content stratified/equal difficulty procedure still has merit by ensuring that the highest and lowest intercorrelations are not significantly different. PM4 also did this, but this method just assigned every third item to a parcel and thus there is no rationale behind the method ensuring the intercorrelations are not significantly different. On the other hand, the method that extremely violated content stratification (PM5) and the method that extremely violated equal difficulty (PM6) generally resulted in a statistically significant difference between the highest and lowest intercorrelations of the multiple-choice parcels. Thus, the content stratified/equal difficulty procedure has merit in ensuring that the intercorrelation range is small so that the parcels are equivalent markers of the test. However, since extreme violations of the rationale behind the content stratified/equal difficulty procedure (i.e., PM5 and PM6) failed to significantly change the fit indicators and thus the conclusion of the confirmatory factor analysis, it appears that approximate content stratification and equal difficulty would be sufficient.

A consistent trend is seen across the grades with respect to the fit of the one-factor model and the fit of the two-factor model. In going from grade 9 to grade 12, the fit of the one-factor model becomes increasingly worse whereas the fit of the two-factor model remains about the same. The tests were examined in more detail to try to come up with a possible explanation for why

the two types of tests appear to be less unidimensional with increasing grade level. The test specifications for the multiple-choice tests (see Appendix A) reveal these tests are somewhat heavier on the life science questions than physical science questions at grades 9 and 10, whereas the reverse is true for grades 11 and 12. At grade 9, with the multiple-choice test heavier on life science and the constructed-response test (Bicycle Science) focusing on physical science, it might be predicted that a one-factor model would yield a bad fit. At grade 12, with the multiple-choice test heavy on physical science and the constructed-response test (Car Power) also focusing on physical science, it might be predicted that a one-factor model would fit fairly well. However, these two predictions contradict the observed trend.

Another possible explanation is that with increasing grade, the content of the constructed-response tests is less related to material actually taught in the classroom, that is, to the material typically found on traditional multiple-choice tests like the TAP. When the tests were examined however, the grade 11 constructed response test (Chemistry Classics) seemed to contain the most items presented in the same way as the material is traditionally taught in high school physical science and chemistry classes. The questions dealing with phase changes, density, and solubility include graphs and tables that are very similar to what is found in most high school texts. On the other hand, the grade 9 constructed-response test (Bicycle Science) tended to present the concepts in ways that are not necessarily encountered in the

classroom. For instance, when students study simple machines, the examples they encounter in lecture and lab activities do not typically include the bicycle and the bones/muscles of the arm examples found on the constructed-response test. Unfortunately, a careful examination of the actual multiple-choice and constructed-response tests failed to provide a viable explanation for why the two types of tests appear to be less unidimensional with increasing grades. Possibly, this trend is spurious.

It should be kept in mind that this study used a convenience sample consisting of subjects who were all from the same school district. Although the descriptive statistics for the multiple-choice test suggests this sample is comparable to the TAP standardization sample, the conclusions of the study would be better supported if a sampling plan for ensuring a representative sample of high school students had been used. In addition to an improved sampling plan, it is desirable that future research use tests from content areas other than science to investigate whether the conclusion about the robustness of this procedure is generalizable across content areas.

CONCLUSION

This study investigated whether a one-factor model or a two-factor model is more appropriate for explaining the covariance between parcels of the TAP multiple-choice and constructed-response science tests. Of particular interest was whether the conclusion about the number of factors is independent of the number of parcels used and how items are assigned to the parcels. Based on inspection of chi-square data and standardized residuals, the two-factor model was generally found to be favored over the one-factor model. This conclusion about the number of factors was consistent across six methods varying the number of parcels and how items were assigned to the parcels. The content stratified/equal difficulty procedure for creating parcels has merit in ensuring that the intercorrelations between parcels from the same test are approximately equal, and thus the parcels are equivalent markers of the test. On the other hand, since the confirmatory factor analysis methodology is fairly robust to violations of the rationale behind the content stratified/equal difficulty procedure, a strict implementation of this procedure is unnecessary.

TABLE A. SUBJECTS AND INSTRUMENTS

GRADE	NUMBER OF SUBJECTS	MULTIPLE-CHOICE TEST TAP, FORM L	CONSTRUCTED-RESPONSE TEST PERFORMANCE ASSESSMENT
9	257	Level 15	Bicycle Science
10	259	Level 16	Biology on Display
11	169	Level 17	Chemistry Classics
12	187	Level 18	Car Power

TABLE B. SUMMARY OF PARCEL METHODS

PARCEL METHOD	NUMBER OF PARCELS	PROCEDURE FOR ASSIGNING ITEMS TO MULTIPLE-CHOICE PARCELS*
PM1	3	content stratified/equal difficulty
PM2	2	content stratified/equal difficulty
PM3	5	content stratified/equal difficulty
PM4	3	every third item
PM5	3	by content (life, earth, physical science parcels)
PM6	3	by difficulty (hard, medium, easy parcels)

*Parcels for the constructed-response tests were held constant; three parcels were created for each constructed-response test using the content stratified/equal difficulty procedure.

TABLE C. SELECTED INDICATORS OF FIT

FIT INDICATOR	INTERPRETATION
CHI-SQUARE VALUE	An overall chi-square value that is not statistically significant indicates the model is a good fit.
CHI-SQUARE PER DEGREE OF FREEDOM RATIO	A value of 2 or lower is typically taken to indicate the model is a good fit (Marsh & Hocevar, 1985).
CHI-SQUARE DIFFERENCE	Used when the two models are nested - when fixing estimated parameter(s) of one model leads to the other model. If difference is not statistically significant, the two models are considered equivalent and the model with fewer estimated parameters is preferred based on parsimony.
STANDARDIZED RESIDUALS	Calculated by dividing fitted covariance minus observed covariance by the asymptotic standard error; interpreted as standard normal deviates. Model is a good fit when few, if any, are $> \|2.58\|$ (Joreskog & Sorbom, 1989, p. 32). Also, positive (underfitting) and negative (overfitting) equally dispersed throughout the matrix of standardized residuals suggests a good fit, as opposed to being confined to only part of the matrix (i.e., a pattern).
MODIFICATION INDICES	For the two-factor model, each parcel has a modification index, which is an estimate of how much the chi-square value would change if the parcel were allowed to load on both factors rather than just one factor. When all are less than 5, a model allowing any one of the parcels to load on both factors would not result in a statistically significant better fit (Marsh & Hocevar, 1985).

TABLE D. DESCRIPTIVE STATISTICS

GRADE 9 (N = 257)

	MULTIPLE-CHOICE TEST		CONSTRUCTED-RESPONSE TEST	
	THIS STUDY	STANDARDIZATION SAMPLE 50 pts	THIS STUDY	32 pts
Mean	26.21	25.73	9.82	
Standard Deviation	9.89	9.61	2.95	
Coefficient Alpha	.900	.894	.503	
Average Difficulty	.5242	.5146	.3069	

Correlation between the two tests = .537

GRADE 10 (N = 259)

	MULTIPLE-CHOICE TEST		CONSTRUCTED-RESPONSE TEST	
	THIS STUDY	STANDARDIZATION SAMPLE 50 pts	THIS STUDY	36 pts
Mean	26.86	26.49	20.20	
Standard Deviation	9.85	10.27	5.47	
Coefficient Alpha	.899	.908	.713	
Average Difficulty	.5372	.5298	.5610	

Correlation between the two tests = .687

GRADE 11 (N = 169)

	MULTIPLE-CHOICE TEST		CONSTRUCTED-RESPONSE TEST	
	THIS STUDY	STANDARDIZATION SAMPLE 50 pts	THIS STUDY	34 pts
Mean	27.49	25.30	14.11	
Standard Deviation	9.85	10.26	4.92	
Coefficient Alpha	.903	.909	.765	
Average Difficulty	.5497	.5060	.4151	

Correlation between the two tests = .644

GRADE 12 (N = 187)

	MULTIPLE-CHOICE TEST		CONSTRUCTED-RESPONSE TEST	
	THIS STUDY	STANDARDIZATION SAMPLE 50 pts	THIS STUDY	33 pts
Mean	27.44	24.66	17.86	
Standard Deviation	9.91	10.81	6.56	
Coefficient Alpha	.905	.920	.822	
Average Difficulty	.5488	.4932	.5412	

Correlation between the two tests = .586

TABLE E. RANGE OF INTERCORRELATIONS

BETWEEN MULTIPLE-CHOICE PARCELS

METHOD	GRADE 9	GRADE 10	GRADE 11	GRADE 12
PM1	.752-.775	.728-.777	.740-.763	.765-.791
PM2	.826	.811	.848	.850
PM3	.589-.716	.609-.715	.596-.699	.621-.708
PM4	.729-.771	.751-.766	.735-.782	.762-.778
PM5	.636-.721*	.594-.745*	.748-.796	.581-.789*
PM6	.608-.737*	.680-.762*	.596-.807*	.560-.788*

BETWEEN CONSTRUCTED-RESPONSE PARCELS

GRADE 9	GRADE 10	GRADE 11	GRADE 12
.246-.325	.454-.510	.518-.618	.555-.649

* difference between lowest and highest intercorrelation is statistically significant, $p<.01$

TABLE F. INITIAL ANALYSIS

GRADE 9 (N = 257)

	ONE FACTOR	TWO FACTORS
CHI-SQUARE	PM1	PM1
df	9	8
chi value	19.27	5.92
p	.023	.657
chi/df	2.141	.740

CHI-SQUARE DIFFERENCE

	PM1
df	1
chi value	13.35
p	<.001

STANDARDIZED RESIDUALS		
	PM1	PM1
> |2.58|	1/15	0/15
pattern	no	no

MISCELLANEOUS	PM1
Correlation between factors	.765
Maximum modification index	1.607

GRADE 10 (N = 259)

	ONE FACTOR	TWO FACTORS
CHI-SQUARE	PM1	PM1
df	9	8
chi value	36.32	7.38
p	.000	.496
chi/df	4.036	.923

CHI-SQUARE DIFFERENCE

	PM1
df	1
chi value	28.94
p	<.001

STANDARDIZED RESIDUALS		
	PM1	PM1
> |2.58|	4/15	0/15
pattern	yes	no

MISCELLANEOUS	PM1
Correlation between factors	.839
Maximum modification index	1.775

TABLE F. INITIAL ANALYSIS (continued)

GRADE 11 (N = 169)

	ONE FACTOR	TWO FACTORS
CHI-SQUARE	PM1	PM1
df	9	8
chi value	54.46	6.62
p	.000	.579
chi/df	6.051	.828

CHI-SQUARE DIFFERENCE

	PM1
df	1
chi value	47.84
p	<.001

STANDARDIZED RESIDUALS	PM1	PM1
> |2.58|	4/15	0/15
pattern	yes	no

MISCELLANEOUS	PM1
Correlation between factors	.762
Maximum modification index	2.510

GRADE 12 (N = 187)

	ONE FACTOR	TWO FACTORS
CHI-SQUARE	PM1	PM1
df	9	8
chi value	109.32	9.87
p	.000	.274
chi/df	12.147	1.234

CHI-SQUARE DIFFERENCE

	PM1
df	1
chi value	99.45
p	<.001

STANDARDIZED RESIDUALS	PM1	PM1
> |2.58|	6/15	0/15
pattern	yes	no

MISCELLANEOUS	PM1
Correlation between factors	.672
Maximum modification index	1.736

TABLE G. VARYING THE NUMBER OF PARCELS

GRADE 9 (N = 257)

ONE FACTOR

CHI-SQUARE	PM1	PM2	PM3
df	9	5	20
chi value	19.27	16.27	37.92
p	.023	.006	.009
chi/df	2.141	3.254	1.896

TWO FACTORS

PM1	PM2	PM3
8	4	19
5.92	3.05	24.45
.657	.550	.180
.740	.763	1.287

CHI-SQUARE DIFFERENCE

	PM1	PM2	PM3
df	1	1	1
chi value	13.35	13.22	13.47
p	<.001	<.001	<.001

STANDARDIZED RESIDUALS	PM1	PM2	PM3
> |2.58|	1/15	3/10	2/28
pattern	no	yes	no

PM1	PM2	PM3
0/15	0/10	0/28
no	no	no

MISCELLANEOUS

	PM1	PM2	PM3
Correlation between factors	.765	.766	.764
Maximum modification index	1.607	1.731	1.704

GRADE 10 (N = 259)

ONE FACTOR

CHI-SQUARE	PM1	PM2	PM3
df	9	5	20
chi value	36.32	33.14	58.69
p	.000	.000	.000
chi/df	4.036	6.628	2.935

TWO FACTORS

PM1	PM2	PM3
8	4	19
7.38	6.17	28.47
.496	.187	.075
.923	1.543	1.498

CHI-SQUARE DIFFERENCE

	PM1	PM2	PM3
df	1	1	1
chi value	28.94	26.97	30.22
p	<.001	<.001	<.001

STANDARDIZED RESIDUALS	PM1	PM2	PM3
> |2.58|	4/15	5/10	5/28
pattern	yes	yes	no

PM1	PM2	PM3
0/15	0/10	1/28
no	no	no

MISCELLANEOUS

	PM1	PM2	PM3
Correlation between factors	.839	.842	.835
Maximum modification index	1.775	1.509	4.211

TABLE G. VARYING THE NUMBER OF PARCELS (continued)

GRADE 11 (N = 169)

	ONE FACTOR			TWO FACTORS		
CHI-SQUARE	PM1	PM2	PM3	PM1	PM2	PM3
df	9	5	20	8	4	19
chi value	54.46	57.28	58.73	6.62	5.12	8.13
p	.000	.000	.000	.579	.275	.985
chi/df	6.051	11.456	2.937	.828	1.280	.428

CHI-SQUARE DIFFERENCE	PM1	PM2	PM3
df	1	1	1
chi value	47.84	52.16	50.60
p	<.001	<.001	<.001

STANDARDIZED RESIDUALS	ONE FACTOR			TWO FACTORS		
	PM1	PM2	PM3	PM1	PM2	PM3
> |2.58|	4/15	6/10	3/28	0/15	0/10	0/28
pattern	yes	yes	yes	no	no	no

MISCELLANEOUS	PM1	PM2	PM3
Correlation between factors	.762	.750	.753
Maximum modification index	2.510	.627	.945

GRADE 12 (N = 187)

	ONE FACTOR			TWO FACTORS		
CHI-SQUARE	PM1	PM2	PM3	PM1	PM2	PM3
df	9	5	20	8	4	19
chi value	109.32	103.05	118.85	9.87	1.75	20.31
p	.000	.000	.000	.274	.782	.376
chi/df	12.147	20.610	5.943	1.234	.438	1.069

CHI-SQUARE DIFFERENCE	PM1	PM2	PM3
df	1	1	1
chi value	99.45	101.3	98.54
p	<.001	<.001	<.001

STANDARDIZED RESIDUALS	ONE FACTOR			TWO FACTORS		
	PM1	PM2	PM3	PM1	PM2	PM3
> |2.58|	6/15	8/10	6/28	0/15	0/10	1/28
pattern	yes	yes	no	no	no	no

MISCELLANEOUS	PM1	PM2	PM3
Correlation between factors	.672	.668	.674
Maximum modification index	1.736	.713	5.230

TABLE H. VARYING HOW ITEMS ARE ASSIGNED TO PARCELS

GRADE 9 (N = 257)

ONE FACTOR

CHI-SQUARE	PM1	PM4	PM5	PM6
df	9	9	9	9
chi value	19.27	23.75	16.63	31.36
p	.023	.005	.055	.000
chi/df	2.141	2.639	1.848	3.484

TWO FACTORS

PM1	PM4	PM5	PM6
8	8	8	8
5.92	10.76	4.46	18.11
.657	.215	.814	.020
.740	1.345	.558	2.264

CHI-SQUARE DIFFERENCE

	PM1	PM4	PM5	PM6
df	1	1	1	1
chi value	13.35	12.99	12.17	13.25
p	<.001	<.001	<.001	<.001

STANDARDIZED RESIDUALS

	PM1	PM4	PM5	PM6
> |2.58|	1/15	2/15	1/15	5/15
pattern	no	no	no	no

PM1	PM4	PM5	PM6
0/15	0/15	0/15	4/15
no	no	no	no

MISCELLANEOUS

	PM1	PM4	PM5	PM6
Correlation between factors	.765	.768	.775	.760
Maximum modification index	1.607	4.855	1.640	9.108

GRADE 10 (N = 259)

ONE FACTOR

CHI-SQUARE	PM1	PM4	PM5	PM6
df	9	9	9	9
chi value	36.32	35.59	30.40	32.23
p	.000	.000	.000	.000
chi/df	4.036	3.954	3.378	3.581

TWO FACTORS

PM1	PM4	PM5	PM6
8	8	8	8
7.38	6.85	4.46	5.62
.496	.552	.813	.690
.923	.856	.558	.703

CHI-SQUARE DIFFERENCE

	PM1	PM4	PM5	PM6
df	1	1	1	1
chi value	28.94	28.74	25.94	26.61
p	<.001	<.001	<.001	<.001

STANDARDIZED RESIDUALS

	PM1	PM4	PM5	PM6
> |2.58|	4/15	2/15	2/15	4/15
pattern	yes	yes	yes	yes

PM1	PM4	PM5	PM6
0/15	0/15	0/15	0/15
no	no	no	no

MISCELLANEOUS

	PM1	PM4	PM5	PM6
Correlation between factors	.839	.839	.842	.842
Maximum modification index	1.775	1.703	1.615	1.654

TABLE H. VARYING HOW ITEMS ARE ASSIGNED TO PARCELS (continued)

GRADE 11 (N = 169)

ONE FACTOR

CHI-SQUARE	PM1	PM4	PM5	PM6
df	9	9	9	9
chi value	54.46	53.92	54.34	55.21
p	.000	.000	.000	.000
chi/df	6.051	5.991	6.038	6.134

TWO FACTORS

PM1	PM4	PM5	PM6
8	8	8	8
6.62	3.76	2.70	4.40
.579	.878	.952	.819
.828	.470	.338	.550

CHI-SQUARE DIFFERENCE

	PM1	PM4	PM5	PM6
df	1	1	1	1
chi value	47.84	50.16	51.64	50.81
p	<.001	<.001	<.001	<.001

STANDARDIZED RESIDUALS

	PM1	PM4	PM5	PM6
> |2.58|	4/15	3/15	3/15	5/15
pattern	yes	yes	yes	yes

PM1	PM4	PM5	PM6
0/15	0/15	0/15	0/15
no	no	no	no

MISCELLANEOUS

	PM1	PM4	PM5	PM6
Correlation between factors	.762	.755	.751	.749
Maximum modification index	2.510	.387	1.164	1.073

GRADE 12 (N = 187)

ONE FACTOR

CHI-SQUARE	PM1	PM4	PM5	PM6
df	9	9	9	9
chi value	109.32	103.49	98.08	109.48
p	.000	.000	.000	.000
chi/df	12.147	11.499	10.898	12.164

TWO FACTORS

PM1	PM4	PM5	PM6
8	8	8	8
9.87	4.94	9.29	14.62
.274	.764	.318	.067
1.234	.618	1.161	1.828

CHI-SQUARE DIFFERENCE

	PM1	PM4	PM5	PM6
df	1	1	1	1
chi value	99.45	98.55	88.79	94.86
p	<.001	<.001	<.001	<.001

STANDARDIZED RESIDUALS

	PM1	PM4	PM5	PM6
> |2.58|	6/15	7/15	8/15	9/15
pattern	yes	yes	yes	no

PM1	PM4	PM5	PM6
0/15	0/15	0/15	1/15
no	no	no	no

MISCELLANEOUS

	PM1	PM4	PM5	PM6
Correlation between factors	.672	.673	.687	.662
Maximum modification index	1.736	1.692	5.193	7.839

APPENDIX A. TEST SPECIFICATIONS FOR MULTIPLE-CHOICE TESTS

CONTENT CLASSIFICATION	NUMBER OF QUESTIONS			
	Grade 9	Grade 10	Grade 11	Grade 12
Nature of Science/Scientific Process Scientific method and inference Analysis of data and information	15	15	14	14
Life Sciences Life processes Characteristics of plants and animals Continuity of life: reproduction, heredity and evolution Environmental interactions; adaptation	15	14	12	11
Earth and Space/Environmental Sciences The earth's surface Atmosphere and weather The universe; the earth in space and motion Forces of nature: constructive and destructive Conservation, renewability, and utilization of the earth's resources	10	10	10	8
Physical Sciences Mechanics, forces, and motion Forms of energy Electricity and magnetism Characteristics and composition of matter Changes and reactions	10	11	14	17
TOTAL	50	50	50	50

APPENDIX B. TEST SPECIFICATIONS FOR CONSTRUCTED-RESPONSE TESTS

CONTENT CLASSIFICATION	NUMBER OF QUESTIONS			
	Grade 9	Grade 10	Grade 11	Grade 12
Nature of Science	8	6	7	8
Science Subject Matter	8	3	9	5
Scientific Concepts and Connections	4	4	4	5
Decision Making/Communication	5	5	6	2
TOTAL	25	18	26	20

REFERENCES

Ackerman, T.A. & Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Bennett, R.E., Rock, D.A. & Wang, M. (1991). Equivalence of free-response and multiple-choice items. Journal of Educational Measurement, 28, 77-92.

Birenbaum, M. & Tatsuoka, K.K. (1987). Open-ended versus multiple-choice response formats - It does make a difference for diagnostic purposes. Applied Psychological Measurement, 11, 385-395.

Cook, L.L., Dorans, N.J., & Eignor, D.R. (1988). An assessment of the dimensionality of three SAT-Verbal test editions. Journal of Educational Statistics, 13, 19-43.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Hinkle, D.E., Wiersma, W., & Jurs, S.G. (1988). Applied statistics for the behavioral sciences (2nd ed.). Boston: Houghton Mifflin.

Joreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. Psychometrika, 36, 109-133.

Joreskog, K.G., & Sorbom, D. (1989). LISREL 7: A guide to the program and applications (2nd ed.). Chicago: SPSS.

Loehlin, J.C. (1992). Latent variable models: An introduction to factor, path and structural analysis (2nd ed.). Hillsdale, N.J.: LEA.

Lord, F.M. (1957). A significance test for the hypothesis that two variables measure the same trait except for errors of measurement. Psychometrika, 22, 207-220.

Marsh, H.W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: First and higher order factor models and their invariance across groups. Psychological Bulletin, 97, 562-582.

Riverside Publishing. (1993a). Performance assessments for TAP & ITED: Manual for scoring and interpretation - Bicycle science, Biology on Display, Chemistry Classics, Car Power (Standardization ed.). Chicago: Author.

Riverside Publishing. (1993b). Tests of Achievement and Proficiency Interpretive guide for teachers and counselors (Forms K and L, Levels 15-18). Chicago: Author.

Traub, R.E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R.E. Bennett and W.C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment (pp. 29-40). Hillsdale, NJ: Lawrence Erlbaum.

Ward, W.C., Frederiksen, N. & Carlson, S.B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Winne, P.H., & Belfry, M.J. (1982). Interpretive problems when correcting for attenuation. Journal of Educational Measurement, 19, 125-134.

TM025815

NCME Annual Meeting, April 9-11, 1996

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Factor Analytic Methods for Determining Whether Multiple-Choice and Constructed-Response Tests Measure the Same Construct	
Author(s): Jim J. Manhart	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☑ ← Sample sticker to be affixed to document

Check here

Permitting microfiche (4"x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

Sample sticker to be affixed to document → ☐

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy.

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Jim Manhart	Position: Graduate Assistant
Printed Name: Jim Manhart	Organization: University of Iowa
Address: 705 Carriage Hill, #1 Iowa City, IA 52246	Telephone Number: (319) 358-7843
	Date: 4/26/96

OVER

CUA

THE CATHOLIC UNIVERSITY OF AMERICA
Department of Education, O'Boyle Hall
Washington, DC 20064
202 319-5120

March 12, 1996

Dear NCME Presenter,

Congratulations on being a presenter at NCME[1]. The ERIC Clearinghouse on Assessment and Evaluation invites you to contribute to the ERIC database by providing us with a written copy of your presentation.

Abstracts of papers accepted by ERIC appear in *Resources in Education (RIE)* and are announced to over 5,000 organizations. The inclusion of your work makes it readily available to other researchers, provides a permanent archive, and enhances the quality of *RIE*. Abstracts of your contribution will be accessible through the printed and electronic versions of *RIE*. The paper will be available through the microfiche collections that are housed at libraries around the world and through the ERIC Document Reproduction Service.

We are gathering all the papers from the NCME Conference. You will be notified if your paper meets ERIC's criteria for inclusion in *RIE*: contribution to education, timeliness, relevance, methodology, effectiveness of presentation, and reproduction quality.

Please sign the Reproduction Release Form on the back of this letter and include it with **two** copies of your paper. The Release Form gives ERIC permission to make and distribute copies of your paper. It does not preclude you from publishing your work. You can drop off the copies of your paper and Reproduction Release Form at the **ERIC booth (23)** or mail to our attention at the address below. Please feel free to copy the form for future or additional submissions.

Mail to: NCME 1996/ERIC Acquisitions
O'Boyle Hall, Room 210
The Catholic University of America
Washington, DC 20064

This year ERIC/AE is making a **Searchable Conference Program** available on the NCME web page (http://www.assessment.iupui.edu/ncme/ncme.html). Check it out!

Sincerely,

Lawrence M. Rudner, Ph.D.
Director, ERIC/AE

[1]If you are an NCME chair or discussant, please save this form for future use.